|  |  |
| --- | --- |
| Seth Gainey | srgainey |
| Kishore Yedavalli | kvyedava |

**Introduction**

The Kepler Mission was launched to search for extra­solar planets (exoplanets) and further the research of planetary formation and the search for ‘earth­like’ qualities such as mass and density, solar transit period and gravitational fields, which may indicate the habitable environment for life outside our solar system. Hundreds of confirmed exoplanets have been discovered in the last five years, due primarily to the data obtained from the Kepler mission. The most recent data for Kepler (up to May 2013) is still under analysis.

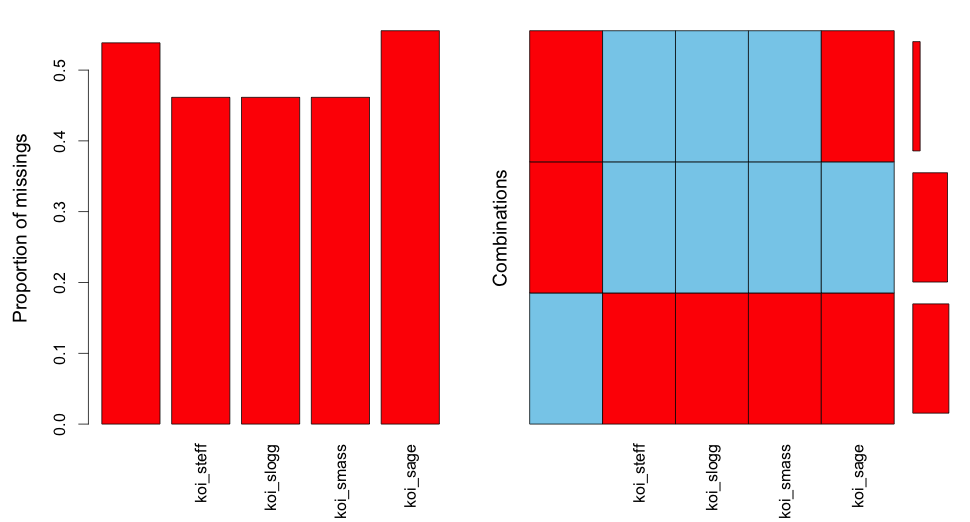
Due to the working status of the dataset, seven new planets confirmations were added just during our analysis (Kepler-77 b, PH2 b, HAT-P-42 b, KOI-142 b, KOI-142 c, OGLE 2012-BLG-358L b, HATS-3 b)

In this project, we examine further the characteristics of those false positives with the intent of seeking underlying relationships that might exist among them. We will examine the criteria used in determining what constitutes a Kepler Object of Interest (KOI) and investigate the attributes of false positives KOIs. Kepler data acquisition is currently in hiatus due to mechanical failure, but data over a period of four years is publicly available (“NASA ­ Kepler Guest Observer Program”).

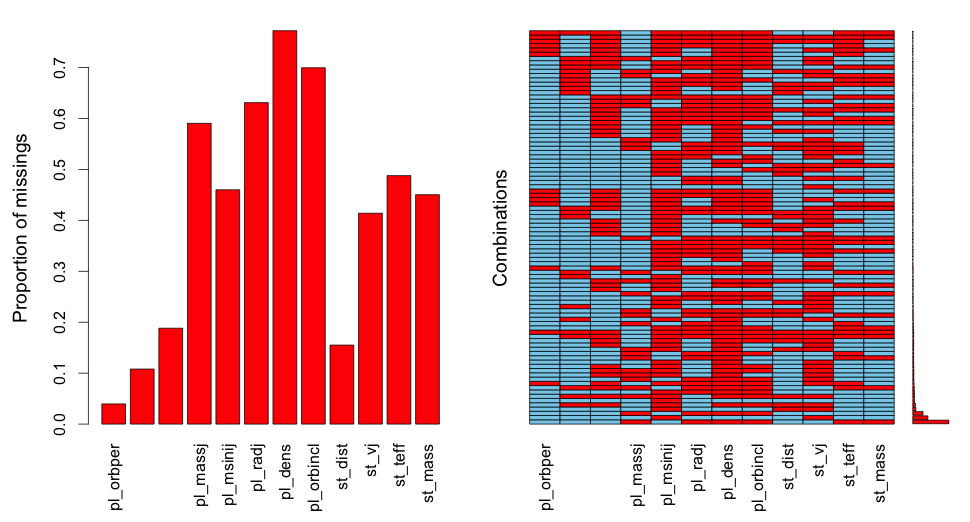
Our analysis covers both the KOI dataset with 2336 rows of 67 attributes and the exoplanet dataset for comparison.

**Implementation**

The Kepler Objects of Interest dataset is downloaded from the Caltech server (<http://exoplanetarchive.ipac.caltech.edu>), and parsed into a dataframe. This is used for preliminary data investigation and a pruned version is passed to Weka to determine attribute importance. There is a high level of missing (NA) data and a high class imbalance. The missing data is often due to provenance: e.g., for exoplanets discovered using radial velocity, it is not feasible to calculate density for that planet.



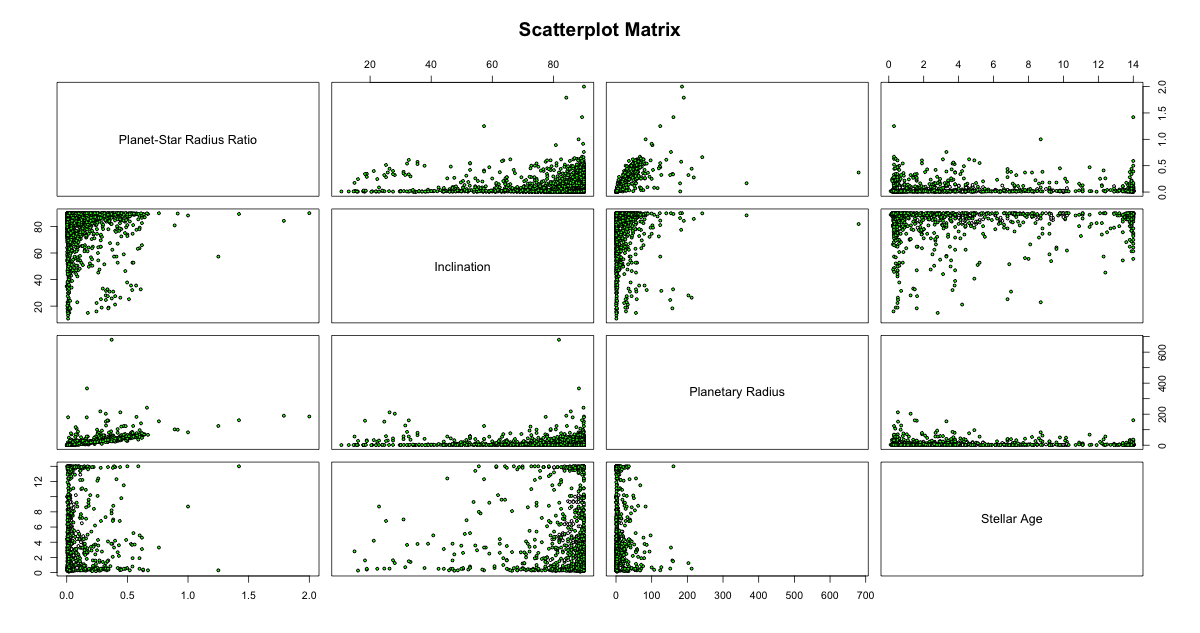
NA Value Visualization for KOIs



NA Value Visualization for Confirmed Exoplanets

|  |  |  |
| --- | --- | --- |
| Class | CONFIRMED | FALSE POSITIVE |
| N | 155 | 2183 |

Class Imbalance

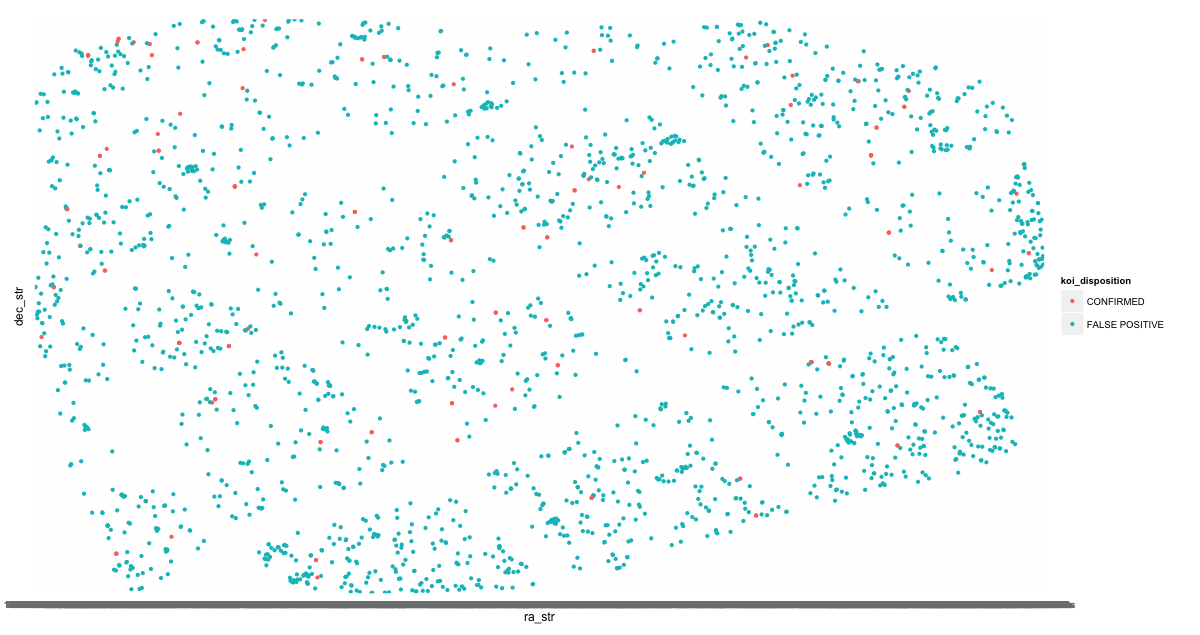
For preliminary investigation, we created a paired scatterplot and a list of attributes ranked by Gain Ratio (Entropy).

Pruned attribute list (from complete list of 67 attributes)

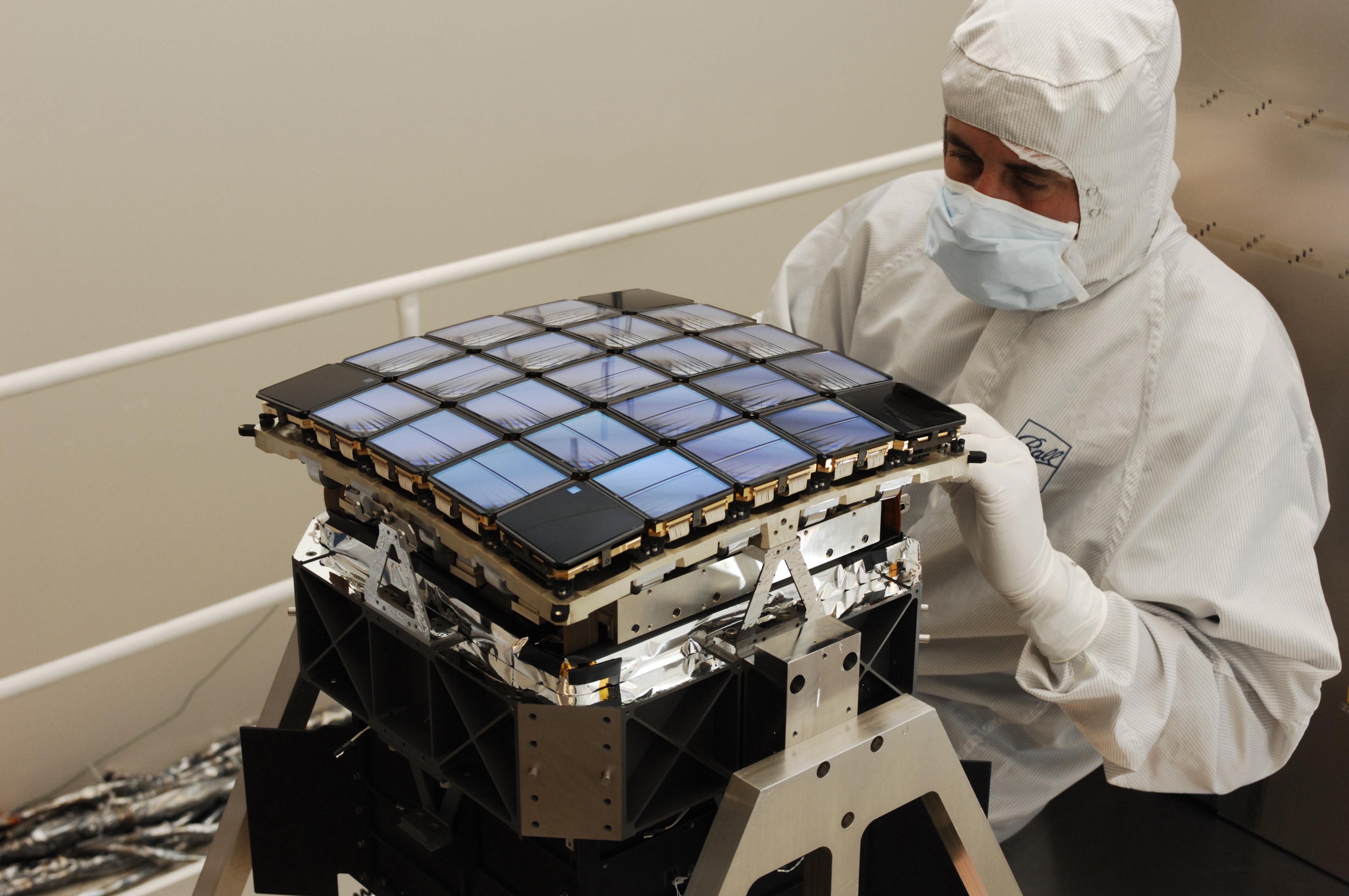
|  |  |  |
| --- | --- | --- |
| average merit | Attribute | Description |
| 0.05 +- 0.001 | koi\_steff | The photospheric temperature of the star |
| 0.048 +- 0.001 | koi\_smass | The mass of the star |
| 0.048 +- 0.001 | koi\_sage | The age of the sta |
| 0.037 +- 0.003 | koi\_time0bk | Time of first detected transit |
| 0.032 +- 0 | koi\_incl | The angle between the plane of the sky (perpendicular to the line of sight) and the orbital plane of the planet candidate |
| 0.031 +- 0.001 | koi\_slogg | The base-10 logarithm of the acceleration due to gravity at the surface of the star |
| 0.03 +- 0.003 | koi\_srad | The photospheric radius of the star |
| 0.027 +- 0.002 | koi\_dor | The distance between the planet and the star at mid-transit divided by the stellar radius. For the case of zero orbital eccentricity, the distance at mid-transit is the semi-major axis of the planetary orbit |
| 0.026 +- 0.002 | koi\_sma | Half of the long axis of the ellipse defining a planet's orbit. For a circular orbit this is the planet-star separation radius. The semi-major axis is derived based on Kepler's third law, i.e., utilizing the orbital period and stellar mass, not scaling the planet-star separation by the stellar radius |
| 0.026 +- 0.001 | koi\_ror | The planet radius divided by the stellar radius |

These attributes purposely exclude data related to provenance or other meta data unrelated to the task of discovering exoplanets without prior knowledge (e.g., date created by the processing pipeline, and location data).

In one case, you can see artifacts of the physical media in the data; right ascension (RA) and declination (DEC) angles are used to measure astronomical locations. This allows us to create a plot of false positive and confirmed planets:



Although there is no location-based pattern of false positive and confirmed planets, we can compare the pattern to the physical charge-couple device (CCD) matrix of the Kepler satellite.



Source: NASA

From the pruned attribute list, we compared several classifiers based on 10-fold cross validation error rate.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Error Rate | RMS Error | TP Rate | FP Rate | Precision | Recall | ROC AUC | Kappa |
| AdaBoost (Logistic Regression Weak Classifier) | 3.2506 | 0.1687 | 0.967 | 0.314 | 0.966 | 0.967 | 0.882 | 0.7134 |
| Bagging | 4.3199 | 0.178 | 0.957 | 0.423 | 0.953 | 0.957 | 0.93 | 0.6049 |
| Naîve Bayes | 26.3045 | 0.48 | 0.737 | 0.091 | 0.939 | 0.737 | 0.916 | 0.2332 |
| RandomForest | 2.7374 | 0.1626 | 0.973 | 0.344 | 0.972 | 0.973 | 0.934 | 0.7399 |
| Simple Logistic Regression | 3.55 | 0.1751 | 0.964 | 0.404 | 0.962 | 0.964 | 0.936 | 0.6615 |

The low error rate and high Kappa statistic of AdaBoost and Random Forest signal that they are well suited to the dataset and not simply exploiting the high class imbalance (i.e, they are actually indicating some signal to the data).

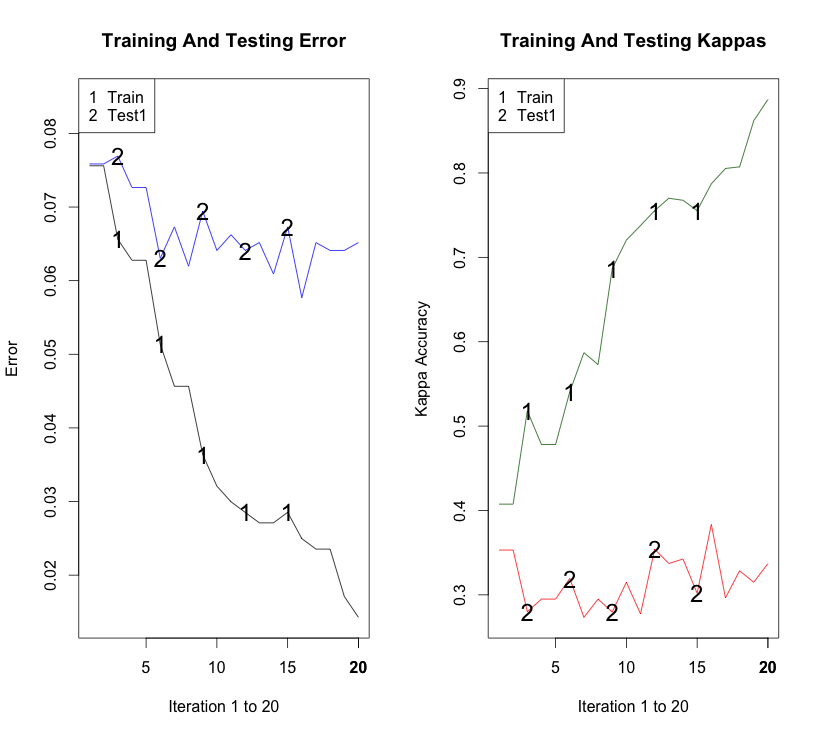
AdaBoost (Logistic Regression) Confusion Matrix

|  |  |  |
| --- | --- | --- |
| a | b | classified as |
| 2159 | 24 | a = FALSE POSITIVE |
| 52 | 103 | b = CONFIRMED |

Random Forest Confusion Matrix

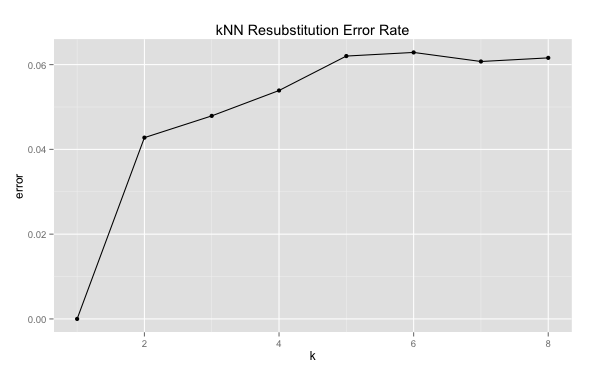
|  |  |  |
| --- | --- | --- |
| a | b | classified as |
| 2176 | 7 | a = FALSE POSITIVE |
| 57 | 98 | b = CONFIRMED |

Both classifiers perform comparably well, but Random Forest has a slightly higher false positive rate. Further investigation of AdaBoost reveals an optimal number of iteration between 10 and 15; any more creates overfitting to the training dataset with no gain in Kappa statistic for the test dataset, and very little reduction in error rate. A 60% partition was used for training and a Logistic Regression weak classifier in this example.

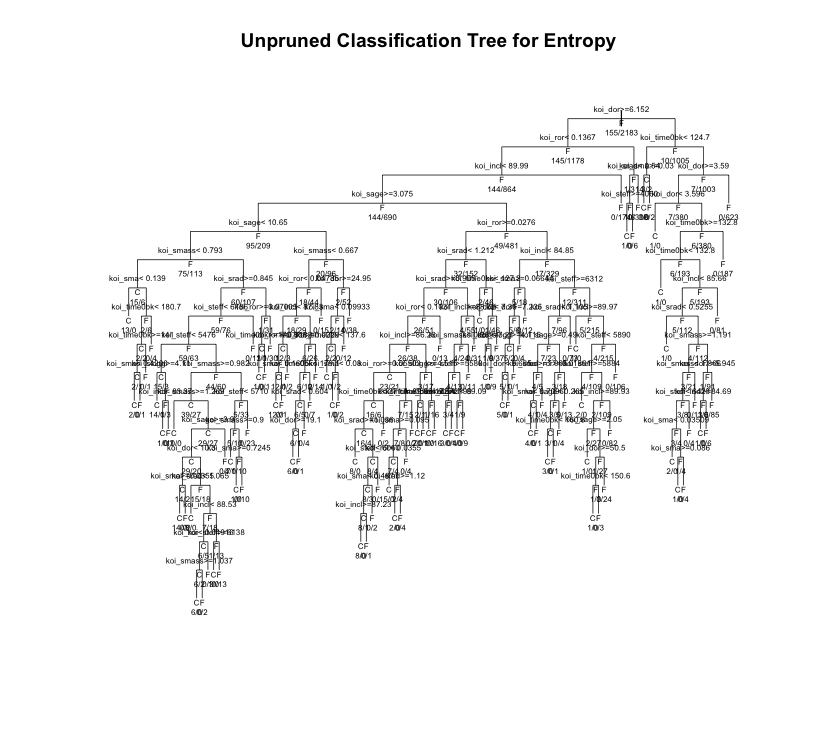


**Failed Approaches**

We also conducted a test of K-nearest neighbor after using imputed values for the NA fields (using the [MICE package](http://cran.r-project.org/web/packages/mice/index.html)). The result was acceptable, although with a higher error rate than AdaBoost or Random Forest). This KNN test was performed on a further reduced subset of attributes (koi\_incl, koi\_sma, koi\_ror, koi\_dor, koi\_srad), but the high-dimensionality and high NA percentage of the dataset makes KNN an unsuitable algorithm.



Decision trees (using rpart and entropy splits) were also investigated, however, the high-dimensionality of the dataset also makes these unsuitable



**Findings**

It was surprising to see how poorly Naïve Bayes suited the dataset, ostensibly due to the dependency between the attributes. And although KNN required substantial cleaning, imputation and pruning of the dataset, the reported error rate was close to the AdaBoost (the lowest error rate). We found that including the Kappa statistic in our analysis was necessary in order to compare classifiers with close error rates, and we would recommend this approach for data sets with a high class imbalance.

**Sources**

M. N. Fanelli, J. M. Jenkins, S. T. Bryson, E. V. Quintana, J. D. Twicken, H. W. Wu, P. Tenenbaum, C. L. Allen, D. A. Caldwell, H. Chandrasekaran, J. L. Christiansen, B. D. Clarke, M. T. Cote, J. L. Dotson, R. Gilliland, F. Girouard, J. P. Gunter, J. Hall, M. R. Haas, K. Ibrahim, K. Kinemuchi, T. Klaus, J. Kolodziejczak, J. Li, P. Machalek, S. D. McCauliff, C. K. Middour, R. Morrris, F. Mullally, S. Seader, J. C. Smith, M. Still, S. E. Thompson, A. K. Uddin, J. Van Cleve, and B. Wohler, Kepler Data Processing Handbook (KSCI­19081­001). http://archive.stsci.edu/kepler/manuals/KSCI­19081­001\_Data\_Processing\_Handbook.pdf

“NASA ­ Kepler Guest Observer Program.” Accessed September 29, 2013. http://keplergo.arc.nasa.gov/ArchiveSchedule.shtml.